# **Day 7 Worksheet - Introduction to counting reads with featureCounts**

Author: Rutendo Sigauke (rutendo.sigauke@colorado.edu)

**Introduction:**

The featureCounts library is part of Subread (written in C) and RSubread (an R wrapper for Subread), and it is a fast tool optimized for counting reads over features (genes, exons, transcripts .etc). To see the full utility of Subreads/Rsubread, see their documentation below:

Subread: <http://subread.sourceforge.net/>

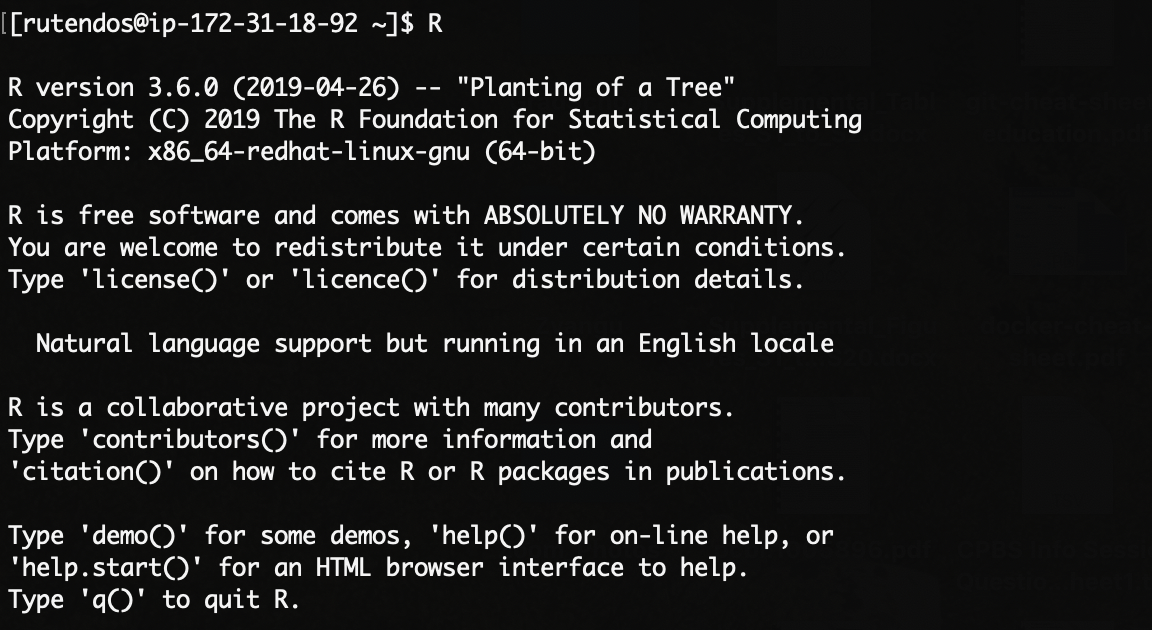
RSubread: <http://subread.sourceforge.net/SubreadUsersGuide.pdf>

Since counting is compute-intensive, this is done on AWS. Usually, we can request multiple threads which makes the counting run faster. **We will be completing the counting section on the server**.

### **Install Rsubread:**

Before running Rsubread, we have to install the library into R. Installation can be done in the R console (shown below).

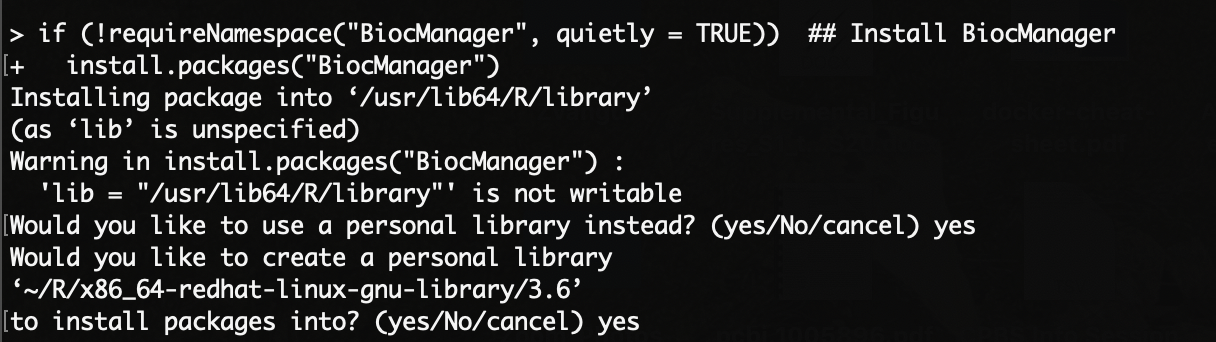
* Type **R** in the terminal



* Rsubread can be found on the **BiocManager**, so to install the counting library, we have to first install **BiocManager**. **BiocManager** library can be installed from the R Comprehensive R Archive Network (CRAN).

**if (!require("BiocManager", quietly = TRUE))**

**install.packages("BiocManager")**

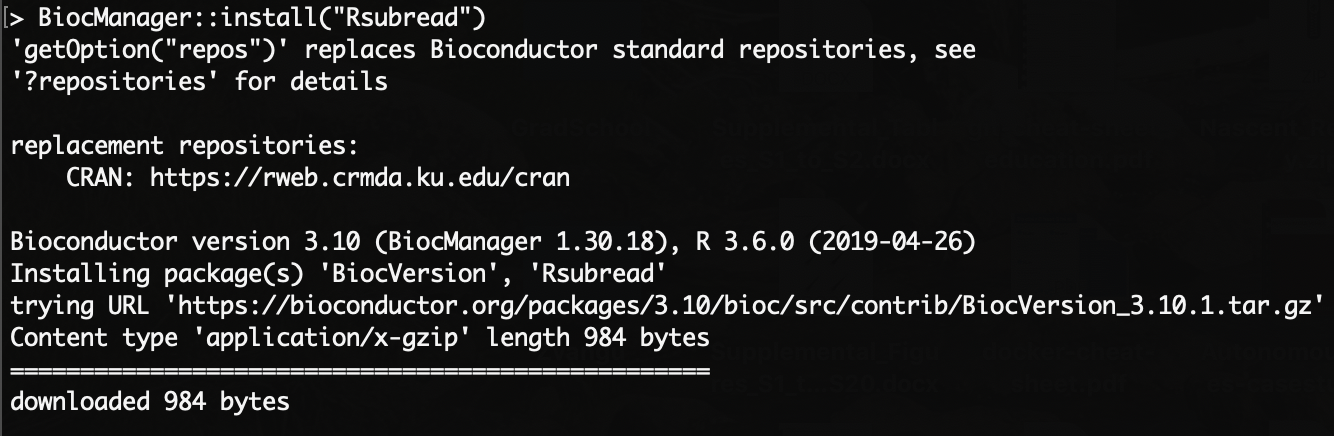


* If you have to set up a personal library, follow the prompt and respond **yes**. The above command will list CRAN mirrors from where to download the packages. Pick one of the US mirriors (e.g. 71).



* Now, we can install **Rsubread** from **BiocManager** to our **R** libraries.

**BiocManager::install("Rsubread")**



**NB:** This will take a few seconds. If the library is installed successfully, it can be loaded as shown below without any errors.



### **Make working directories:**

Make the necessary working directories for running featureCounts.

1. Use the command **pwd** to determine what directory you are in and if necessary, **cd** to the directory that you want to place your new featureCounts directory in.

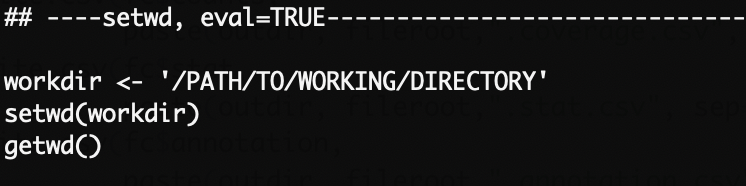
2. Make a new directory in using the **mkdir** command. Use command **ls -lsh** to confirm the folders are present.



3. Edit **featureCounts** scripts using **vim <script>**. This will open the scripts in the text editor.

**Edit R script:**

* Set your working directory

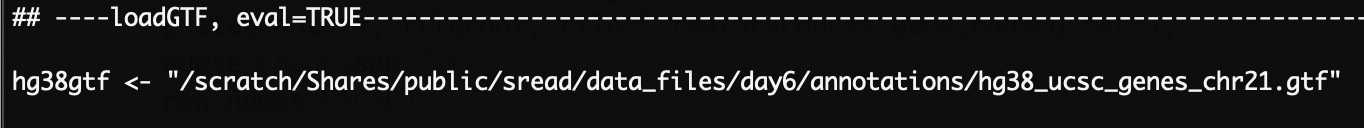


**NB:** The output folders will be generated based on your **workdir**

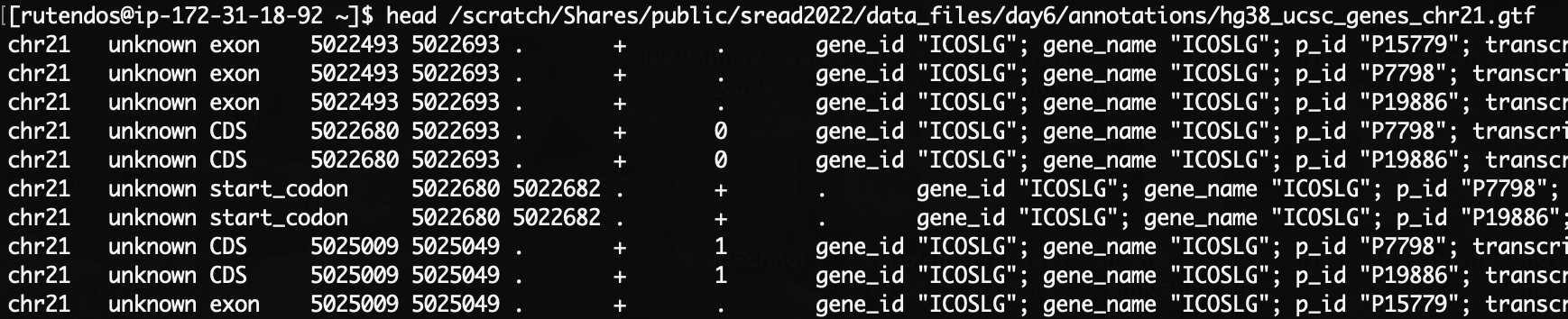
* Create a variable for the bam folder.



* Check what annotations are being used, and make sure the path is correct.



**NB:** Take a look at the GTF file structure in the commandline (exit vim or R console). Note all the different features represented for each feature. Also, you will see that the file has several columns, with the **first** column is the chromosome ID, the **second** column is the name of the source from which the feature was derived (eg. RefSeq, Ensembl, UCSC or HAVANA). The **third** column is the label for the feature (e.g. exon, CDS, start\_codon). This field is used by featureCounts to determine the features to to count reads over. The **fourth** and the **fifth** columns are start and end coordinates respectively. The **sixth** column is the score of the feature, the **seventh** the strand, the **eighth** is phase for CDS features (If phase=0, the codon begin at the first base of CDS nucleotide; if phase=1 the codon begin at the second base of CDS nucleotide; if phase=2 the codon begin at the third base of CDS nucleotide.). Lastly, the **nineth** column contains additional feature annotations.



**Edit sbatch script:**

* Edit then sbatch script including the **SBATCH headers** and path to the **d7\_featureCounts.R** script.



* Run the sbatch script!

5. You can find the summary of counts in the output counts folder. There are five different files:

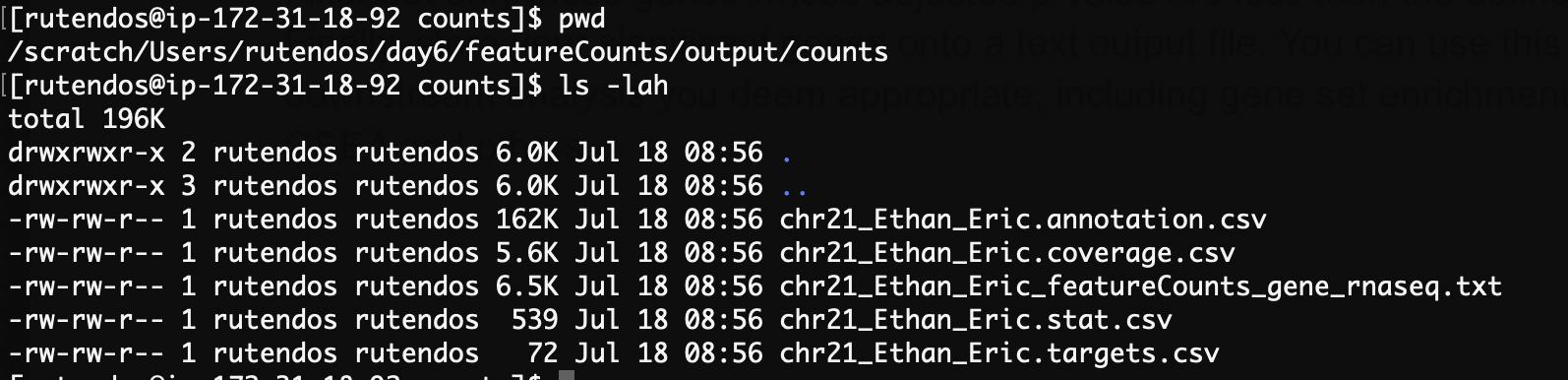
- **featureCounts\_gene\_rnaseq.txt** : GeneID, Length, Counts

- **coverage.csv** : Counts

- **.stat.csv** : Coverage Statistics

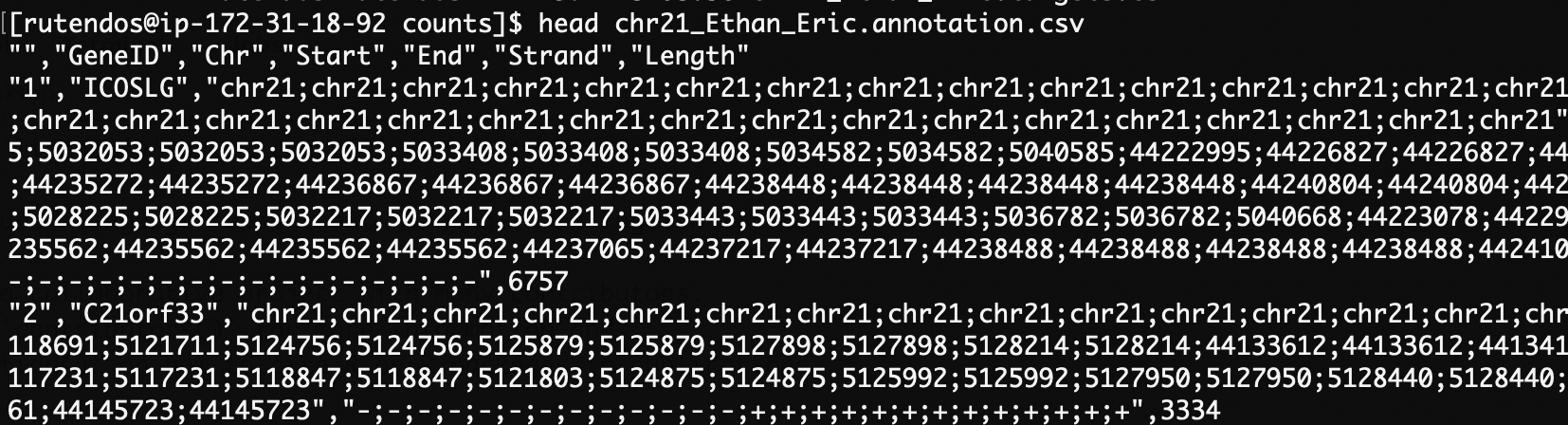
- **.annotation.csv** : GeneID, Chromosome, Start, End, Strand, Length

- **.targets.csv** : file names for input `bam` files

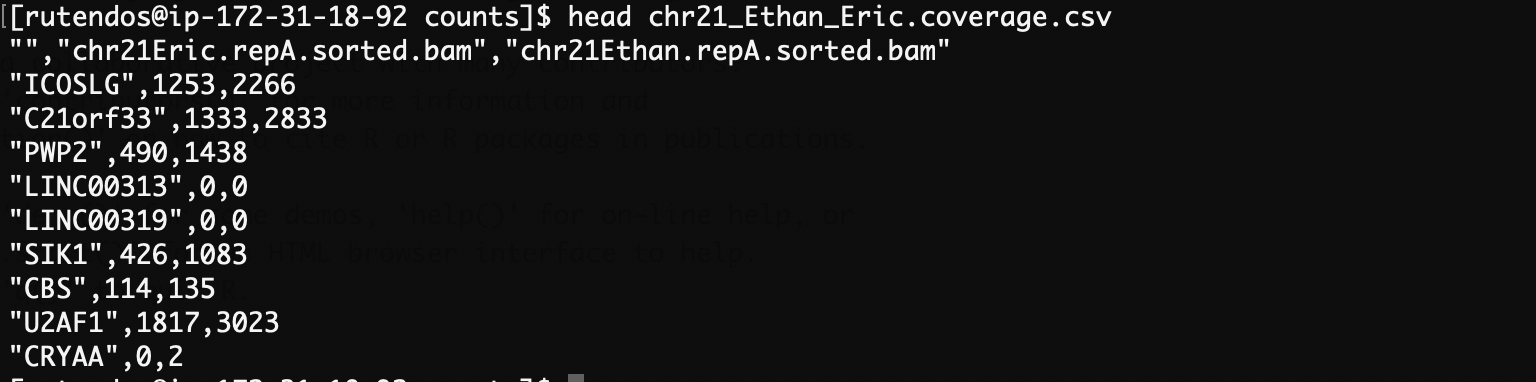


6. Open and explore each of the files in the terminal ( with **head** or **less**). You can also move them to your local computer using **scp** or **rsync**, and load the files in RStudio.

- **chr21\_Ethan\_Eric.annotation.csv**



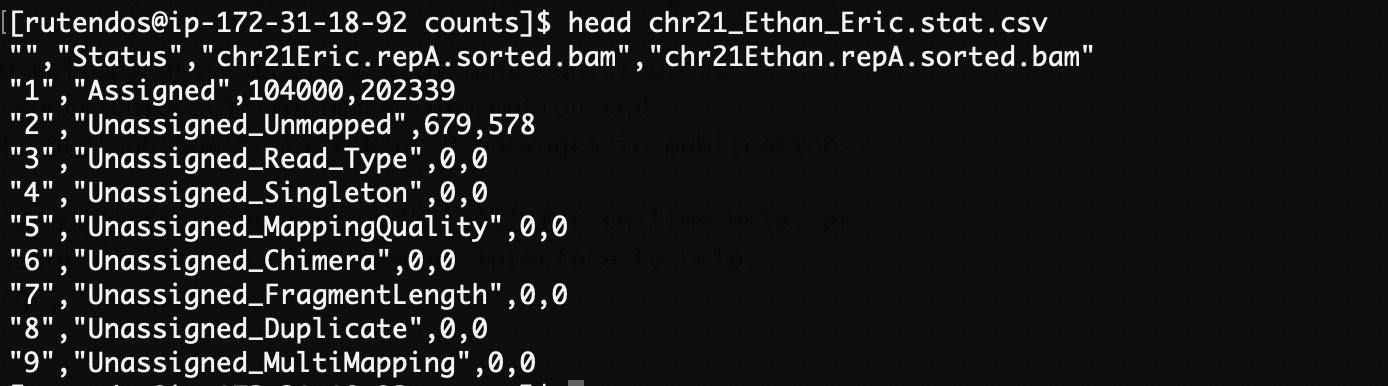
- **chr21\_Ethan\_Eric.coverage.csv**



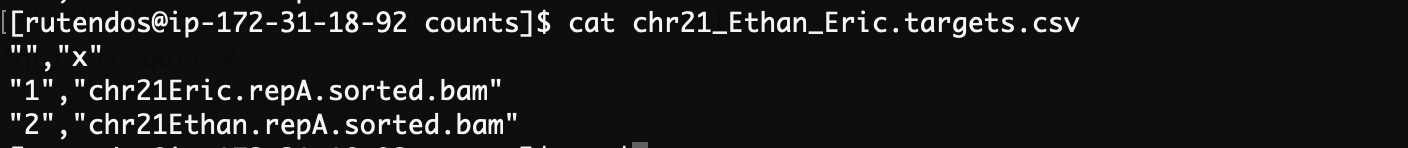
**- chr21\_Ethan\_Eric\_featureCounts\_gene\_rnaseq.txt**



- **chr21\_Ethan\_Eric.stat.csv**

****

- **chr21\_Ethan\_Eric.targets.csv**

****